ISSN 2616-5791 Vol.5, Issue 9: 40-45, DOI: 10.25236/AJMHS.2024.050907

Research on the Strategy of MedKGGPT Model in Improving the Interpretability and Security of Large Language Models in the Medical Field

Jinzhu Yang*

AI Research, Dyania Health Inc, Jersey City, New Jersey, 07310, United States
jinzhu.yang0625@yahoo.com
**Corresponding author*

Abstract: *With the deep application of artificial intelligence technology in the medical field, especially the widespread deployment of large-scale language models, improving their interpretability and security has become an urgent and important issue to be solved. This article focuses on medical diagnostic tasks and proposes an innovative MedKGGPT model strategy aimed at significantly enhancing the interpretability and security of medical decisions by integrating machine learning and knowledge reasoning methods. The MedKGGPT model achieves a deep integration of medical expertise and data-driven learning through a carefully designed architecture, including knowledge inference modules, machine learning modules, and intelligent fusion modules. This model first constructs a detailed ontology knowledge and business rule library for medical diagnosis, providing a solid theoretical foundation for subsequent reasoning and decision-making. Subsequently, the machine learning module learns and extracts key features from massive medical data, while the knowledge reasoning module utilizes expert knowledge to analyze and validate these features. The intelligent fusion module is responsible for efficiently integrating the two results, constructing a decision evidence chain through credibility evaluation, ensuring that the model maintains high classification accuracy while also having a high degree of interpretability in its decision-making process. In order to verify the effectiveness of the MedKGGPT model, this paper selected cervical cancer cell recognition based on liquid based cytology examination images as an experimental case. The experimental results show that the model not only performs well in classification accuracy, but also continuously optimizes performance during the iteration process. More importantly, it can provide clear and explicit explanatory paths for each medical decision, greatly enhancing the trust of medical personnel in artificial intelligence assisted diagnosis. The contribution of this article is not only in proposing a novel MedKGGPT model strategy, enriching the interpretability theory of artificial intelligence models in the medical field, but also providing practical and feasible solutions for improving the security of medical decision-making and promoting trust between doctors and patients. This research achievement has profound significance for promoting the healthy development of artificial intelligence technology in the medical field.*

Keywords: *MedKGGPT Model, Medical Field, Interpretability, Security, Integration of Machine Learning and Knowledge Reasoning*

1. Introduction

With the rapid development of artificial intelligence technology, especially significant breakthroughs in natural language processing (NLP) and deep learning, large-scale language models (LLMs) such as the GPT series are increasingly being applied in the medical field. These models, with their powerful text generation, understanding, and reasoning abilities, have shown great potential in assisting medical decision-making, medical record recording, health counseling, and other areas. However, despite the broad application prospects of LLMs in the medical field, the black box nature of their decision-making process, namely the lack of interpretability and transparency, has become a major obstacle to their widespread application and increased trust.In the medical field, any decision is directly related to the patient's life safety and health well-being, so the decision-making basis and reasoning process of the model must have a high degree of interpretability and transparency. Currently, although most LLMs can generate high-quality text output, their internal mechanisms are complex and it is difficult to directly explain the reasons and basis for generating specific outputs. This opacity not

ISSN 2616-5791 Vol.5, Issue 9: 40-45, DOI: 10.25236/AJMHS.2024.050907

only makes it difficult for doctors and patients to trust the results of the model, but may also lead to serious consequences such as misdiagnosis and missed diagnosis.In order to enhance the interpretability and security of large-scale language models in the medical field, researchers have begun to explore various strategies, including but not limited to the integration of knowledge graphs, the development of interpretable methods, and the construction of security mechanisms. Among them, integrating knowledge graphs with LLMs has become a promising solution. Knowledge graph, as a structured knowledge representation method, can organize and represent professional knowledge in the medical field in the form of a graph, providing rich background knowledge and reasoning basis for LLMs. By combining knowledge graphs with LLMs, the interpretability and accuracy of the model can be improved to a certain extent. Based on this background, this article proposes the MedKGGPT model, aiming to improve the interpretability and security of LLMs in the medical field by integrating knowledge graphs and large-scale language models. The MedKGGPT model not only inherits the powerful language generation and understanding capabilities of the GPT series models, but also provides rich medical knowledge and reasoning paths for the model by integrating knowledge graphs in the medical field. This fusion strategy enables the MedKGGPT model to infer and make judgments based on medical knowledge when generating medical related texts, thereby improving the interpretability and accuracy of its decisions.This article first analyzes the challenges faced by current LLMs in the medical field in terms of interpretability and safety, and elaborates on the necessity of integrating knowledge graphs with LLMs. Subsequently, the design concept, implementation methods, and technical details of the MedKGGPT model were introduced in detail, including interpretability theory, machine learning models used, knowledge reasoning methods, and the combination strategy of machine learning and logical reasoning. On this basis, a series of experiments and case studies were conducted to verify the effectiveness of the MedKGGPT model in improving the interpretability and safety of LLMs in the medical field.Through the research in this article, we hope to provide a new solution for the interpretability and security issues of LLMs in the medical field, and promote the widespread application and in-depth development of artificial intelligence technology [1] in the medical field. At the same time, we also hope that the research results of this article can provide useful references and insights for the interpretability and security research of large-scale language models in other fields.

2. Correlation Theory

2.1. Overview of Interpretability

Interpretability [2] is crucial in artificial intelligence systems, referring to the ability of the system to present its decision-making process and results in terms that are understandable to humans. This feature serves as a bridge between humans and complex algorithms, enabling people to perceive and trust the behavior and predictions of models. At present, interpretability research is mainly divided into two categories: intrinsic interpretability and post interpretability. Essentially interpretable models, such as naive Bayes and decision trees, have simple and transparent structures that are easy to understand, but often have limited accuracy and are difficult to apply to high-precision fields. On the contrary, high-precision complex models such as deep learning models, although having excellent performance, often lack self interpretability due to their "black box" characteristics. To balance these two aspects, researchers have proposed compromise solutions such as generalized additive models. The post interpretability method aims to provide explanations for complex models through techniques such as individual condition maps, automatic rule extraction, permutation feature importance, and model distillation after model training. These methods each have their own emphasis, but they all aim to improve the comprehensibility and trustworthiness of the model. In the field of medical diagnosis, it is particularly important to combine the ideas of intrinsic interpretability and post interpretability, as directly linking the pathological causes of specific symptoms for explanation is more likely to gain the recognition of doctors than simply exploring algorithmic mechanisms. This article proposes an interpretable intelligent model that integrates machine learning and knowledge reasoning, aiming to process medical image data through convolutional neural networks such as VGGNet, and combine knowledge reasoning modules to provide high-performance and interpretable intelligent decision support for medical diagnosis.

2.2. Ontology Knowledge

Ontology plays a core role in knowledge expression and sharing in the fields of computer science

ISSN 2616-5791 Vol.5, Issue 9: 40-45, DOI: 10.25236/AJMHS.2024.050907

and artificial intelligence, especially in semantic web technology. It originated from philosophy, but in computer science, it specifically refers to a clear specification of conceptual models, used to describe and share clear, unique, and universally recognized knowledge systems within a specific field. Domain ontology [3] as an application form of ontology, focuses on the formal description of entities, concepts, relationships, and rules in specific fields (such as medical diagnosis). By constructing principles such as clarity, consistency, and scalability, it ensures the effective organization and reuse of domain knowledge. Network Ontology Language (OWL), as a standard language for describing ontology, was developed by W3C and provides powerful expressive power to support machine interpretability of web content. In terms of knowledge reasoning, the axioms and facts in the OWL ontology are expressed through the Semantic Web Rule Language (SWRL), achieving a semantic based reasoning mechanism. This mechanism enables interpretable intelligent models in the field of medical diagnosis to systematically organize medical knowledge and provide accurate and interpretable diagnostic support through intelligent reasoning.

2.3. Combining Machine Learning with Logical Reasoning

The integration of machine learning and logical [4] reasoning has become a key path to address the challenges of complex problems. Although both have achieved significant technological progress, a single technology often finds it difficult to fully meet the complex needs of the real world. Machine learning excels at extracting patterns and patterns from data, but has limitations in logical reasoning ability; And logical reasoning, as the core of intelligent behavior, is crucial for understanding complex relationships and making reasonable explanations. Therefore, the organic combination of machine learning and logical reasoning has become a cutting-edge hot topic in current research, indicating a new trend in solving complex problem models. Existing research has achieved this fusion through various methods, such as using declarative first-order logical rules to enhance neural network performance, combining logical induction with statistical induction to optimize learning tasks, and developing plug and play modules such as relational networks to solve the problem of relational inference. These studies not only demonstrate the potential of combining the advantages of both, but also explore novel models such as induced learning, achieving collaborative work between neural perception and symbol reasoning. On this basis, the paper further explores the method of using knowledge reasoning to assist machine learning results [5] in logical interpretation, and introduces an evolutionary mechanism to support continuous optimization, aiming to promote the development of this field to a deeper level.

3. Research Method

3.1. Explainability Model

This chapter elaborates on an interpretability model proposed for the field of medical diagnosis that integrates machine learning and knowledge reasoning. The design inspiration for this model comes from the process of human decision-making by combining perception and reasoning. When physicians determine whether cervical cells are diseased, they first obtain an overall impression through perception, then infer detailed features based on professional knowledge, and finally combine the two to draw conclusions. The model consists of a knowledge inference module, a machine learning module, and a knowledge inference fusion machine learning module. The knowledge reasoning module provides a domain ontology library and a rule library, laying the foundation for reasoning and decision-making; The machine learning module extracts feature information from the data through target feature classifiers and multiple sub feature classifiers. The knowledge reasoning fusion machine learning module is the core of the model, which extracts the correlation features (sub features) between expert knowledge and data, uses the results of sub feature classifiers for knowledge reasoning, and combines the target feature classification results to make the final decision through an evolutionary method based on credibility evaluation. The evolution process not only achieves the interpretability of the model, but also continuously improves the classification performance of the model through iterative optimization. The entire process simulates the thinking of human decision-making, ensuring the effectiveness and reliability of the model in complex medical diagnostic tasks.

3.2. A Fusion Method for Credibility Assessment

This chapter elaborates in detail on the fusion method based on credibility evaluation [6], which is the core of achieving model interpretability and improving classification performance. Firstly, the key

ISSN 2616-5791 Vol.5, Issue 9: 40-45, DOI: 10.25236/AJMHS.2024.050907

data structure of the result evidence chain is defined, which records the parameter values of the classifier and rule library used by the model in obtaining the target feature result Rc and inference result Rr in the form of a directed acyclic graph. By constructing machine learning result evidence chains (including target feature results and sub feature results evidence chains) and inference result evidence chains, it ensures that the model can trace the specific cause and optimize it in case of result errors. Subsequently, reliability calculation methods for Rc and Rr were proposed, which comprehensively considered multiple factors such as dataset quality, network model quality, probability values of classifier observation results, and reliability of rule libraries, ensuring the comprehensiveness and accuracy of credibility. Finally, a detailed description of the evolutionary decision-making process was provided, which simulated the human decision-making process by combining perception and reasoning by comparing the credibility of Rc and Rr with the threshold a, and made corresponding decisions based on different situations, including explaining the results, optimizing the internal structure of the model, or trusting oneself more. In the case where Rc and Rr are the same and both reliable, the model achieves post hoc interpretability; The existence of the result evidence chain ensures the essential interpretability of the model in the event of failure, enabling the model to gradually improve classification performance through continuous iteration. Through the detailed discussion in this chapter, a solid foundation has been laid for verifying the effectiveness and superiority of the model through specific case studies in the future

3.3. MedKGGPT Model

The MedKGGPT model provides an innovative and effective strategy for improving the interpretability and security of Large Language Models (LLMs) [7] in the medical field. This model not only improves the classification performance of the model by integrating machine learning and knowledge reasoning techniques, but also significantly enhances the transparency and interpretability of its decision-making process, thereby meeting the needs of high security and high trust in the medical field.

The core of the MedKGGPT model lies in its structural design, which consists of a knowledge inference module, a machine learning module, and a knowledge inference fusion machine learning module. The knowledge reasoning module provides the model with rich medical knowledge and logical reasoning capabilities by constructing domain ontology libraries and rule libraries; The machine learning module utilizes deep learning techniques to learn feature representations of cell images from a large amount of data; The fusion module cleverly combines the two and achieves intelligent fusion of target feature results and inference results through the method of result evolution, improving the classification accuracy and interpretability of the model.

In terms of interpretability, the MedKGGPT model introduces the concept of result evidence chain, which details the important parameters of the classifier and rule library used by the model in obtaining classification results. This design allows the model to trace back to the specific cause when the result fails, and improves the performance of the model by optimizing the relevant parts. At the same time, when two results (target feature result and inference result) are the same and reliable, the model can provide a detailed explanation path, including sub feature classification results and rules used, thereby enhancing the trust and understanding of doctors in model decision-making.

In terms of security [8], the MedKGGPT model ensures high accuracy, robustness, and stability when processing medical data through a rigorous model validation and evaluation system. In addition, the model also has the ability to self optimize and iterate, and its performance will gradually improve as the amount of classified data increases. This design enables the model to adapt to datasets of different scales and continuously improve in practical applications.

4. Results and Discussion

4.1. Experimental Model

In the field of cervical cancer screening, the MedKGGPT model deeply integrates medical knowledge graphst [9] oembed key abnormal cervical epithelial cells and their morphological features such as LSIL, HSIL, SCC, ASC-US, ASC-H into the model, and is closely related to the eight sub feature (f1-f8) system of clear cervical squamous epithelial cell images segmented from TCT slices provided by the hospital. The cervical squamous epithelial abnormal cell recognition ontology library O, constructed using OWL language and Prot é g é platform, not only defines relevant medical classes

ISSN 2616-5791 Vol.5, Issue 9: 40-45, DOI: 10.25236/AJMHS.2024.050907

and their relationships, but also enhances the model's ontology based reasoning ability, enabling it to generate diagnostic reports containing detailed medical explanations. At the same time, the MedKGGPT model establishes a real-time monitoring and feedback loop mechanism, continuously optimizes algorithms to improve diagnostic accuracy and interpretability, and strictly complies with privacy protection and compliance requirements to ensure the security of medical data and respect for patient privacy, bringing a more intelligent, interpretable, and secure diagnostic solution to the field of cervical cancer screening.

4.2. Experimental Analysis

In the strategy research of applying the MedKGGPT model to improve the interpretability and security of large-scale language models in the medical field, we constructed a complete domain ontology, rule set, and classifier group for the specific case of ASC-H cervical abnormal cell recognition. By conducting classification experiments on a validation set containing 400 annotated images of cervical squamous epithelial cells, we validated the superiority of the MedKGGPT model in three aspects. Firstly, the model effectively integrates target feature classifiers and knowledge reasoning methods that support sub feature results through evolutionary methods, achieving mutual correction between the two, resulting in significantly higher overall classification accuracy than using either classifier alone. This is clearly demonstrated by comparing the accuracy of individual classifiers with the evolved model. Secondly, the MedKGGPT model demonstrates strong interpretability, demonstrating through specific cases how to transform the sub feature results of cell images into entity relationships in the ontology library, and using the Drools inference engine for rule reasoning, ultimately generating pathological explanations that are easy for doctors to understand. Finally, as the classification process continues, the model gradually improves classification accuracy through iterative evolution, and this trend is supported by the steady improvement of the model's accuracy in the increasing number of classification cells. In summary, the application of the MedKGGPT model in the field of cervical cancer screening not only improves the accuracy and safety of classification, but also significantly enhances the interpretability of medical diagnosis, providing strong support for the development of medical intelligence [10-11].

5. Conclusion

With the rapid development of artificial intelligence technology, especially its widespread application in the field of medical diagnosis, the interpretability and security issues of large-scale language models are becoming increasingly prominent. Therefore, this article proposes an innovative MedKGGPT model that successfully solves the problem of traditional models lacking transparency and trust by deeply integrating machine learning and knowledge reasoning techniques. The MedKGGPT model not only includes a knowledge reasoning module and a machine learning module, but also creatively introduces a knowledge reasoning fusion machine learning module. By extracting key sub features, constructing a detailed domain ontology and rule library, and training an efficient classifier group, it achieves accurate recognition and interpretable decision-making of medical image data. Experimental verification shows that the model not only improves classification accuracy, but also significantly enhances the interpretability of decision results, providing a more transparent and reliable intelligent solution for medical diagnosis. Looking ahead to the future, the MedKGGPT model is expected to make more breakthroughs in introducing probabilistic reasoning, enhancing the interpretability of sub feature classifiers, expanding knowledge bases, exploring multimodal fusion, and strengthening security and privacy protection, further promoting the deep application and development of artificial intelligence in the medical field.

References

[1] Cao, Y, Cao, P, Chen, H, Kochendorfer, K. M, Trotter, A. B, Galanter, W. L,...& Iyer, R. K. (2022). Predicting ICU admissions for hospitalized COVID-19 patients with a factor graph-based model. In Multimodal AI in healthcare: A paradigm shift in health intelligence (pp. 245-256). Cham: Springer International Publishing.

[2] Giménez, Maite, Fabregat-Hernández, Ares, Fabra-Boluda, Raül, et al. A Fine-Grained Study of Interpretability of Convolutional Neural Networks for Text Classification. International Conference on Hybrid Artificial Intelligence Systems. Springer, Cham, 2022. DOI:10. 1007/978-3-031-15471-3_23.

[3] Al-Aswadi F N, Chan H, Hoon G K, et al. Enhancing relevant concepts extraction for ontology

ISSN 2616-5791 Vol.5, Issue 9: 40-45, DOI: 10.25236/AJMHS.2024.050907

learning using domain time relevance. Inf. Process. Manag. 2023, 60: 103140. DOI: 10. 1016/j. ipm. 2022. 103140.

[4] De Abreu Araújo, Ivo, Hidaka Torres R, Neto N C S. A Review ofFramework forMachine Learning Interpretability. International Conference on Human-Computer Interaction. Springer, Cham, 2022. DOI: 10. 1007/978-3-031-05457-0_21.

[5] Chen, H, Yang, Y, & Shao, C. (2021). Multi-task learning for data-efficient spatiotemporal modeling of tool surface progression in ultrasonic metal welding. Journal of Manufacturing Systems, 58, 306-315.

[6] Xiao H, Yang P, Gao X W M. Basic uncertainty information hesitant fuzzy multi-attribute decision-making method with credibility. Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology, 2023, 45 (5): 8429-8440.

[7] Lykov, Artem, and D. Tsetserukou. "LLM-BRAIn: AI-driven Fast Generation of Robot Behaviour Tree based on Large Language Model." ArXiv abs/2305. 19352 (2023).

[8] Li F, Wang S. Secure Watermark for Deep Neural Networks with Multi-task Learning. 2021. DOI: 10. 48550/arXiv. 2103. 10021.

[9] Wu X, Duan J, Pan Y, et al. Medical Knowledge Graph: Data Sources, Construction, Reasoning, and Applications. Big Data Mining and Analytics, 2023, 6 (2): 201-217. DOI: 10. 26599/BDMA. 2022. 9020021.

[10] Cadario R, Longoni C, Morewedge C K. Understanding, explaining, and utilizing medical artificial intelligence. Nature Human Behaviour, 2021. DOI: 10. 1038/s41562-021-01146-0.

[11] Cao Y, Cao P, Chen H, Kochendorfer K. M, Trotter A. B, Galanter W. L, & Iyer R. K. Predicting ICU admissions for hospitalized COVID-19 patients with a factor graph-based model. In Multimodal AI in healthcare: A paradigm shift in health intelligence. Cham: Springer International Publishing. 2022, 245-256